Mansoura University
Faculty of Engineering
Mansoura Engineering Journal

A New Intrusion Detection Strategy Based on Combined Feature Selection Methodology and Machine Learning Technique

Shereen H. Ali

KEYWORDS:
Intrusion detection system, Feature Selection, Machine Learning, Genetic Algorithm.

***Abstract*— Intrusion detection system is a significant security mechanism that monitors network traffic to assist prevents unwanted access to network resources. Effective intrusion detection is an important issue for defending networks against potential intrusions. In this paper, a new intrusion detection strategy is proposed. The recommended intrusion detection strategy is divided into three steps: (i) Preparing step, (ii) Feature selection step, and (iii) Classification step. Preparing step gathers and analyzes network traffic in readiness for training and testing. Feature selection step aims to choose the significant features for detecting intrusion attacks form preparing step. It comprises of two successive feature selection modules, which are; quick selection module and precise selection module. Precise selection module deploys genetic algorithm as a wrapper method, whereas quick selection module relies on filter. Based on the most effective features identified by feature selection step, the classification step seeks to detect intrusion attacks with the least amount of time penalty. It contains two phases: prioritized naive bayes phase and distance encouragement phase, which avoids the problems of naive bayes classifiers. The presented intrusion detection strategy beats other previous approaches using the NSL-KDD dataset, according to the experimental tests. Intrusion detection strategy provides the highest accuracy, precision, recall and F1-measure with values equal to 97.6%, 98.24%, 98.14%, and 98.11% respectively with minimum time penalty.**

I. INTRODUCTION

UNAUTHORIZED attacks on computers and networks are detected using an intrusion detection system [1,2]. If an intrusion is detected, alarms have been observed to be emitted by these systems. Based on the detection mechanism, intrusion detection systems can be divided into two groups [3]; (i) Signature-based, which compares specific patterns found in the network, such as bytes' sequences, to a signature database already in existence. (ii) Anomaly-based, which compares a network behavior to a known baseline, and it is good at detecting both known and new intrusions. It is critical to identify intrusion attacks quickly and accurately in order to avoid infection of network resources. Machine learning has recently become a popular research tool [4]. Machine learning is thought to be a useful method for detecting intrusion attacks [5, 6]. Several machine learning approaches for intrusion detection have been introduced. Unfortunately, they have a number of flaws, including (i) limited detection accuracy, and (ii) extended

Received: (02 October, 2021) - Revised: (10 November, 2021) - Accepted: (11 November, 2021)

Corresponding Author: *Shereen H. Ali, Assistant Professor of Communications & Electronics Engineering Department, Delta Higher Institute for Engineering & Technology, Mansoura, Egypt, (e-mail: engshereen2011@gmail.com).*

detection. Naïve Bayes (NB) classifier is a straightforward machine learning algorithm that is highly robust [7, 8]. Regardless of the fact that it has a simplistic design and relies on simplified principles, NB has performed admirably in a variety of challenging real-world scenarios, including disease prediction, text classification, and traffic risk management [9, 10, 11]. Nevertheless, given the target value, the unreasonable claim that those features are autonomous and evenly legitimate [7], in some cases, the effectiveness of NB may be inferior. To overcome this obstacle, several approaches have been suggested, involving feature selection and prioritization [7, 8]. This study introduces an effective Intrusion Detection Strategy (IDS) for identifying intrusion attacks. IDS consists of three cascaded steps, which are; (i) Preparing Step (PS), (ii) Feature Selection Step (FSS) and Classification Step (CS). During PS, the network traffic is obtained and analyzed to equip data for training and testing. During FSS, the most significant features for detecting intrusion attacks form PS has been selected by employing a proposed a Combined Feature Selection Methodology (CFSM). CFSM integrates filter and wrapper approaches, and is divided into two modules: Quick Selection Module (QFM), which employs numerous filter methods, and Precise Selection Module (PSM), which utilizes Genetic Algorithm (GA) as a wrapper method. Actually, filter techniques can give quick selection, but they lack accuracy since; feature dependencies are ignored, and the judgment must be made just once. Wrapper techniques, such as GA, can mitigate for filter method flaws by providing precise selection by taking into account feature relationships and the connection with the deployed classifier. Nevertheless, when compared to filter methods, it cannot deliver fast selection. As a result, CFSM is able to pick the effective features because it evolves filter techniques for quick selection, wrapper methods to provide precise selection, and it take into account feature correlations and connections with the classifier. The CS employs a novel classification technique to provide quick and precise intrusion detection depending on the features picked. The proposed classification approach focuses on improving efficiency as well as resolves the shortcomings of NB through: specifying priorities to the chosen features, resulting in a Prioritized Naive Bayes (PNB) then regulating PNB judgment using distance among the item being categorized and a middle of predefined categories. As a result, the suggested classification system is comprised of two phases: (a) the Prioritized Naïve Bayes Phase (PNBP), where the PNB classifier makes a primary judgment about the item's relevance to all of the predefined categories, (b) the Distance Encouragement Phase (DEP), where last judgment will be taken. Recent intrusion detection methodologies were compared to the suggested IDS. According to the findings of the experiments, IDS exceeds other alternatives because it delivered the highest detection efficiency. The paper is organized as follows: In Sect. II, the related work about intrusion detection techniques is reviewed. Section III presents a detailed explanation of the proposed Intrusion Detection Strategy. In Sect. IV, the experiments are presented and the results are analyzed. In Sect. V, the paper is concluded.

II. RELATED WORK

In [12], a two-phase intrusion detection system was presented. The logistic regression (LR) algorithm has been employed in conjunction with GA method in the first phase of the feature selection approach. Otherwise, the artificial neural network (ANN) approach has been used for classification in the second phase. Several evolutionary-based approaches, including the particle swarm optimization (PSO) method, were used to train the ANN. The performance of the offered frameworks was validated using the NSL-KDD dataset. The PSO-ANN obtained an accuracy of 88.90% and was trained in 74 seconds, according to the results. The GA-ANN, on the other hand, had an accuracy of 83.11% and was trained in 134 seconds.

In [13], the NSL-KDD dataset has been used to construct a bidirectional long short-term memory (BLSTM) approach in conjunction with an attention mechanism (BAT-MC) for feature extraction. The most significant features necessary for an optimal classification approach were captured using the attention algorithm. The BAT and the BAT-MC were both subjected to the experimental procedures. The accuracy for the BAT and BAT-MC were 82.56% and 84.25%, respectively, according to the data.

In [14], an adaptive synthetic sampling (ADASYN) technology was combined with a convolutional neural network (CNN) to create an intrusion detection system. The ADASYN approach was first employed to lower the sensitivity of the algorithm to any type of class imbalance. Second, the split convolution module is where the CNN algorithm used in this study came from the split-based (SPC-CNN) model. During the training phase, the SPC-CNN was utilized to limit the impact of undesired information. Finally, the modeling procedure is carried out using the AS-CNN algorithm in conjunction with ADASYN and SPC-CNN. The suggested framework performance was assessed using the NSL-KDD dataset. In addition, the authors used the RNN and CNN as baseline models. The RNN had a detection accuracy of 69.73 percent, according to the results. The CNN attained a phenomenal accuracy of 80%, whereas the AS-CNN achieved a score of 68.66%.

In [15], a wrapper-based attribute selection technique based on the differential evolution (DE) algorithm was implemented for intrusion detection. The extreme learning machine (ELM) classifier was employed to evaluate the specified feature sets in this study. The DE-ELM was put to

the test on the NSL-KDD dataset. The DE-ELM achieved an accuracy of 80.15% for binary classification configurations, according to the experimental data.

III. THE PROPOSED INTRUSION DETECTION STRATEGY (IDS)

The proposed IDS is depicted in Fig. 1. The proposed strategy is performed through three steps, which are: (i) Preparing Step (PS), (ii) Feature Selection Step (FSS), and (iii) Classification Step (CS). The three steps of the proposed IDS will be discussed in detail in the following sections.

A. Preparing Step (PS)

PS collects and processes network traffic to prepare it for use during training and testing. Any packet filtering tool can be used to collect the required dataset. The data realized is stored upon a log file or a database. The data is therefore submitted to further evaluation before being used in the subsequent steps. Data evaluation consists of three stages: (i) redundant instances are erased from the dataset using data mitigation, (ii) adjustment of attack types, this assigns each attack type to one of the main attack classes, and (iii) the capable of changing non-numeric data items into a consistent numeric form is known as data normalization.

B. Feature Selection Step (FSS)

This section will provide a successful methodology known as Combined Feature Selection Methodology (CFSM) for selecting the significant collection of features that can describe Intrusions attacks. The CFSM is a combination approach that involves filter and wrapper techniques. It is divided into two modules: (i) the Quick Selection Module (QSM), which employs numerous filter methods, and (ii) the Precise Selection Module (PSM), which employs the GA. On a same dataset, different filter techniques will be applied individually in QSM so that each technique can swiftly select a distinct

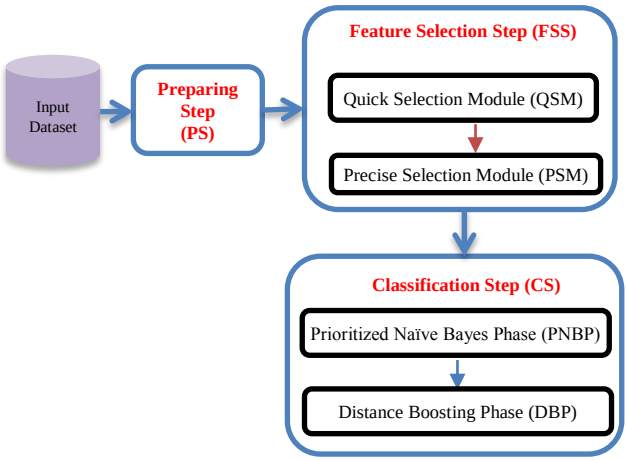

Fig. 1. The recommended IDS

Subset of features. The outputs of the filter techniques will be utilized as a starting population for GA in PSM to accurately choose the significant collection of features. Last, the appropriate feature selection will improve the intrusion detection quality of the model. To construct CFSM, suppose there is Feature Map={s_1, s_2,....,s_d}. Moreover, the input data for learning 'k' objects represented by A={R_1, R_2,...., R_k} as well as testing data about the 'v' objects are represented by V= {U_1, U_2,....,U_v}. Individual object of $R_i \in A$ & $U_j \in V$ is represented by a set of 'd' features that are arranged in a specific sequence; R_i (s_1, s_2, s_3, ..., s_d) = [s_{1i}, s_{2i}, s_{3i}, . . . , sdi] & E_j (s_1, s_2, s_3, ..., s_d) = [s_{1j}, s_{2j}, s_{3j}, . . . , s_{dj}]. Thus, every object R_i & U_j represented in 'd' dimension set of features. The following steps of the CFSM approach employing 'b' filter methods are shown in Fig. 2. First, after performing the PS, the input dataset should be provided to QSM for concurrent implementation of the 'b' filter techniques. The outputs of QSM will then be submitted to PSM, which will produce the GA's starting population. The number of chromosomes in the original population equals 'b, as shown in Fig. 2. Furthermore, the chromosomal values in QSM are the outcome of filter algorithms. Then, until a termination condition is met, GA cycles will be run. And at last, the perfect chromosome achieves the desired group of features, which should be assessed using a NB [16]. GA formulates initial population, which is a collection of chromosomes (O). Each chromosome represents features like a bit vector with the same size as the number of features in the input dataset. Chromosome bits can have a value = '0' or '1'. While a value of '0' in the chromosome's j-th status represents that the j-th feature is not present in the subset, a value of '1' represents that the feature is exist [17, 18]. Selection, crossover, and mutation are three physiologically inspired GA operators that are employed to generate a new cycle of chromosomes [17]. A perfect chromosome is chosen by the selection operator. The crossover operator mixes best chromosomes in order to create better children in the next cycle. A chromosome is changed locally by a mutation operator in order to create an effective one. Finally, as illustrated in Fig.2, GA implements sequential operations. The genetic evaluation (fitness) function in PSM indicates the efficiency of the NB classifier, to choose the best described features from the input dataset. The chromosome with the highest fitness value is the best. There are 'b' chromosomes in the GA (O) initial population, which include the outcomes of 'b' filter methods in QSM as starting solutions. Evaluating the efficiency of the NB classifier should produce the fitness value of each chromosome. The three GA operators should then be executed. The probability of selection value (Pros) is allocated to the chromosomes in O throughout the selection process in order to choose the best two parents. The probability of the crossover value (Proc) is specified both for parents in the crossover operation to indicate whether or

not the crossover process would be completed among them to develop new children in the following cycle. The probability of mutation value (Prom) is given to each child in the mutation process to reflect whether the mutation operation would be executed on each child or not. The phases of the selection operation will be replicated till the new population is the same size as the original. The algorithm will then be terminated by looking at the number of cycles. If there are more cycles than the number of cycles, the former stages from the assessment step will be replayed; otherwise, the chromosomes in the population would be assessed as final results using only the assessment step. Lastly, the optimum collection of characteristics represents the chromosome with the desired fitness value. In Fig.3 CFSM algorithm is depicted. The parameters used in CFSM algorithm is depicted in Table II.

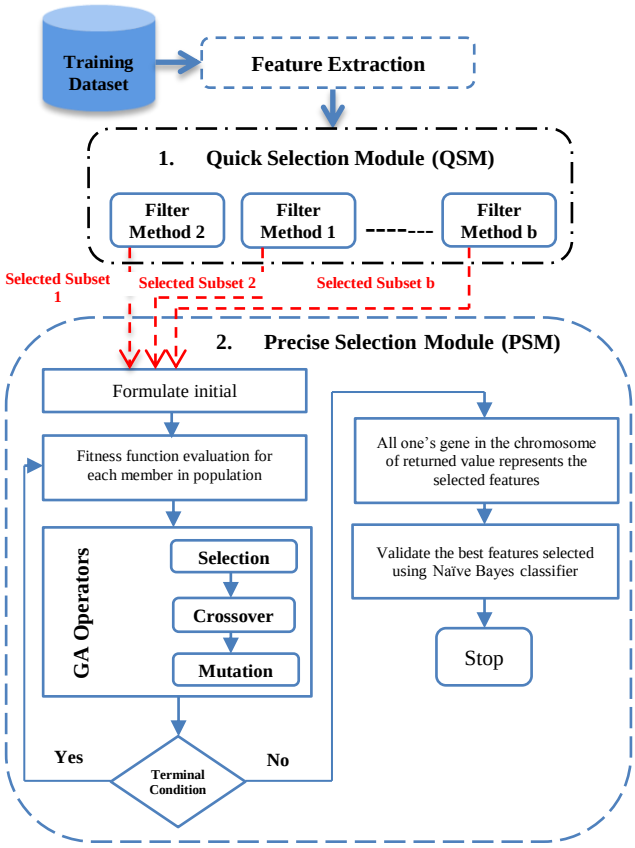

Fig. 2. The steps that followed in CFSM

To better understand the concept, consider the following four QSM filter methods: Information Gain (IG-STD) [19], ImpCHI [20], Fisher score (Fi) [21], and Reversed Correlation Algorithm (RCA) [22]. Furthermore, imagine that input dataset has six features (d=6); Feature Map = {s_1, s_2, s_3, s_4, s_5, s_6}. Following the application of IG-STD, ImpCHI, Fi, and RCA to the input dataset, it is presumed that those approaches result in the collection of gathered features; {s_1, s_3, s_5, s_6}, {s_3, s_4, s_6}, {s_1, s_2, s_3, s_4, s_6}, and {s_1, s_2, s_5, s_6}. As a result, in PSM initial population (O), these 4 groups of features are used as 4 chromosomes (M_1, M_2, M_3, M_4). The GA is then implemented based on a number of hypotheses stated in Table I. According to the hypotheses stated in Table I, GA is applied in two cycles, resulting in a new population with new values at four chromosomes: M_1= {0,1,1,1,1,0}, M_2={1,1,0,1,1,1}, M_3={0,0,0,0,1,1}, and M_4={1,1,1,0,0,1}. After examining M_1, M_2, M_3, and M_4, it is determined that M4 has the highest fitness value, indicating that M_4 is the best chromosome for a given group of traits. Finally, in the input dataset, the most affected features are; {s_1, s_2, s_5, and s_6}.

C. Classification Step (CS)

As previously mentioned, NB classifier faces a shortage of performance as a result of the typical belief that all features are extremely relevant and autonomous. To address this flaw, two problems should be addressed in order to ensure highest efficiency and mitigate for the reliable classifier's limitations: (a) awarding priorities to the selected input dataset features, thus, the conclusion is PNB, (b) PNB's judgment can be influenced via utilizing distance-based influencing. The distance among the item being categorized itself and middle of the predefined categories determines this influencing [23]. This CS consists of phases; (i) Prioritized Naive Bayes Phase (PNBP), and (ii) Distance Encouragement Phase (DEP). During PNBP, a PNB classifier is used to make primary judgments of the input item's relevance to each of the categories being considered. The PNB classifier makes a judgment based on a feature priority vector that is generated. Otherwise, the DEP will make the ultimate judgment based on the item relevancy level projected via PNBP. As a result, the fresh item gets swiftly sorted into one of the predetermined categories under consideration. The details of both phases of the classification step are covered in the following sections.

a) Prioritized Naïve Bayes Phase (PNBP)

In this section, the input item to just be classified is appointed a relevancy degree based on a PNB classifier for each class label. Because efficiency is critical in intrusion detection systems, each selected feature is then prioritized based on the effectiveness of a NB classifier. The priority of the feature x_z, defined as p_z, is a feature influence indicator that measures the proportion of model efficiency lost when x_z is removed out feature collection. Feature prioritization is an important task that can improve detection efficiency. A feature's priority is determined by its positive impact on the overall system efficiency. The feature priority can be calculated as in (1).

TABLE I
THE HYPOTHESIS UNDERLYING THE USE OF GA IN PSM

Hypothesis	Value
Number of cycle to perform	2
Size of Population	4
Pro_s	Random [0,1]
Pro_c	0.9
Pro_m	0.1
M	6

TABLE II
PARAMETERS USED IN CFSM ALGORITHM

TDS	**Training data set contents of training objects and its features, TDS=(D,FM).**
D	Training objects.
TED	Testing data set contents of testing objects and its features, TED=(D,FM).
FM	Features of training or testing object, FM=$x_1,\ldots\ldots x_d$.
H	Chromosome x with highest accuracy value.
Pro_s	Probability of selection.
Pro_m	Probability of mutation.
Pro_c	Probability of crossover.
O	Initial population.
t	Probability distribution.
t(q)	Probability distribution value of member q.
e(q)	Fitness value of chromosome q.
M	Group of chromosomes in population,M=$M_1\ldots.M_{nc}$
M'	Group of new chromosomes in next generation of population; M'=$M'_1\ldots.M'_{nc}$.
O'	Next cycle of population.
n_c	No. of chromosomes in population "population size" that equals to No. of filter methods; n_c=g.
subset (y)	New Input dataset with n_R items.
t	No. of features in training and testing data set, t=|FM|.
b	No. of filter methods in FS^2.
Q	Testing objects

$$p_z = \text{efficiency}\,(+\,x_z) - \text{efficiency}\,(-x_z) \quad (1)$$

Where p_z represents the beneficial effect of feature x_z , *efficiency(* $+\ x_z$ *)* represents system's efficiency since x_z involved in the feature collection, and *efficiency* $(-x_z$ *)* represents system's efficiency since x_z dismissed. The normalized priority for individual feature evaluated via (2).

$$Np_z = \frac{p_z}{\max_{\forall p_d} p_d} \quad (2)$$

A feature priority list is created, which records the normalized priority of every feature selected during the FS^2. The Relevancy Degree (RD) of the reference item T_x among the category c_i is computed via (3) [7].

$$RD\,(T_x,c_i) = P(c_i) * \prod^{d}_{\substack{j=1 \\ where\ p_j \in R^+}} P\left(x_j \middle| c_i\right)^{N_{p_j}} \quad (3)$$

Where *RD* $(T_x$ *,*c_i *)* represents relevance degree of T_x *to* considered a category middle c_i , *P(*c_i *)* represents prior probability of category c_i , Np_j represents normalized priority for x_j, *P(*x_j *|*c_i *)* represents conditional probability of x_j considered a class c_i .

b) Distance Encouragement Phase (DEP)

Before a final judgment would be reached, an item is being categorized into those established categories. For achieve this goal, all items are first anticipated into the d-dimension FM under investigation. In a FM, the middle of that individual category can be achieved by involving e examples in d-dimension FM via (4).

$$Mid = \left\{\frac{\sum_{q=1}^{e} A_q^1}{e}, \frac{\sum_{q=1}^{e} A_q^2}{e}, \ldots., \frac{\sum_{q=1}^{e} A_q^d}{e}\right\} \quad (4)$$

Where *Mid* represents category middle of regarded d-dimension feature map, *e* represents the amount of examples in the category, as well as A_q^i reflects a q-th example's i-th dimension valuation. Then, the reference's item Association Score (AS) for each predefined category can be calculated via (5).

$$AS\,(T_x,c_i) = \frac{RD(T_x,c_i)}{Dis(T_x,Mid\langle c_i\rangle)} \quad (5)$$

Combined Feature Selection Methodology (CFSM) Algorithm

- ***Input:***
  - *TDS= (D,FM); Training dataset.*
  - *TED= (Q,FM); Testing dataset.*
  - *d=|FM|; No. of features in training and testing dataset.*
  - *b= number of filter techniques in FS^2.*
  - *Pro_s =prob. of selection.*
  - *Pro_c =Prob. of crossover.*
  - *Pro_m =prob. of mutation.*
- ***Output:***
  - *H= chromosome M with highest accuracy value.*
- ***Steps:***

// implementing 'b' filter methods on training and testing dataset.

1: **For every filter method y** ∈ b.
2: Determine the subset of selected features for every method as *subset (y)*.
3: **End For**

// construct initial population of GA.

4: Put 'b' Subsets as the values of 'n_c' chromosomes in an initial population (P) with chromosomes denoted by (M).

// calculate fitness value of each chromosome.

5: Calculate an accuracy of the employed classifier as an evaluation function for each chromosome M ∈ O.

// applying selection method using "Roulette wheel".

6: Define a probability distribution (t) over the members of (O) where t(M) ≠ t(M).

7: Select two chromosomes M_i , M_j according to t, Pro_s; where I, j ∈ nc, I ≠ j.

// applying crossover method using "single point crossover".

8: Apply crossover to M_i and M_j to produce new offsprings M_i' and M_j' according to Pro_c.

// applying mutation method using "flip bit mutation".

9: Apply mutation to M_i' and M_j' with respect to Pro_m.

10: Insert M_i' and M_j' to O'(the next cycle).

11: **If (no. of chromosomes in O' less than O) Then**

12: Go to step 7.

13: **Else**

14: Let O ← O'; replace chromosomes value in O with O'.

15: **End If**

16: **If (there are more generations to process) Then**
17: Go to step5.
18: **Else**
19: Return M that contains highest value of e(q) in H, where all one' genes in this chromosome represent the selected features.
20: **End If**

Fig. 3. CFSM algorithm

Where *Dis* (T_x ,*Mid* $<c_i>$) considers the Euclidian distance among T_x and a c_i middle. Determining a distance among both h_x & h_y in the *d*-dimension feature map as depicted in Fig. 4 regarding three- predefined classes is determined via (6):

$$\text{Dis}\,(h_x, h_y) = \sqrt{\sum_{i=1}^{d}(h_x^i - h_y^i)^2} \qquad (6)$$

Last, the predefined category of the T_x , named as *Target*(T_x) , is determined via (7) [7].

Target (T_x) = $argmax_{ci\in C}$[AS (T_x,c_i)]

Target (T_x) = $argmax_{ci\in C}[\frac{RD(T_x,c_i)}{Dis(T_x,Mid\langle c_i\rangle)}]$

$$\text{Target } (T_x) = argmax_{ci\in C}\left[\frac{P(c_i)*\prod^{d}_{j=1} P(x_{j|c_i})^{Np_j}}{Dis(T_x,Mid\langle c_i\rangle)}\right] \quad \text{where } p_{j\in R^+} \qquad (7)$$

IV. Experimental Results

The proposed Intrusion Detection Strategy (IDS) will be examined in this subsection. IDS is divided in three steps, as follow: Preparing Step (PS), Feature Selection Step (FSS), and Classification Step (CS). In PS, network traffic is gathered and analyzed to organize the data for use during training and testing. CFSM is a feature selection approach described in FS2 for picking useful features from input datasets. Then, during the feature prioritization step, those features are prioritized using the NB to award priority for individual recognized feature given the impact upon categorization efficiency, ensuring that each feature is significant. To allow the PNB classifier to make the first judgment, those priority features were used. The separation among the checked item and the category middle in DEP is then used to make the final judgment. Four performance metrics: accuracy, precision, recall, and the f1-score are employed to estimate each aspect of the suggested strategy in the following sections [24].

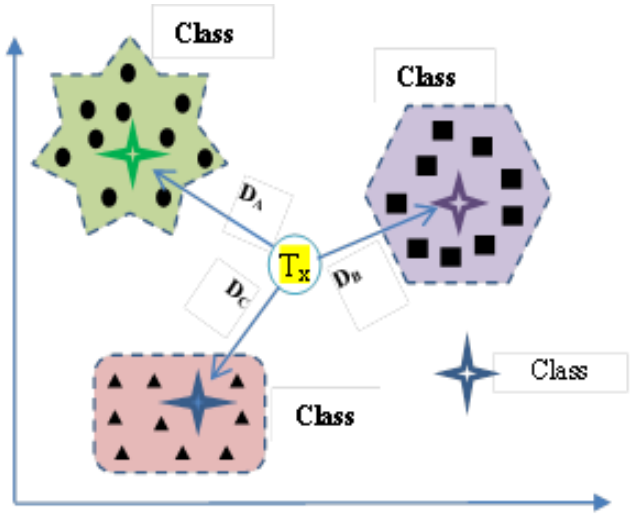

Fig.4. Calculation of the distance to class center

A. Dataset Description

The NSL-KDD dataset [25, 26], was used to implement the recommended IDS as well as the considered rivals. The NSL-KDD dataset is an improved version of the KDD Cup '99 dataset [25]. The NSLKDD dataset is proposed as a solution to some of the concerns with the KDD Cup '99 dataset. The training dataset in the KDD Cup '99 dataset contains 4,94,021 patterns, whereas the testing dataset contains 3,11,029 patterns. The training dataset for the NSL-KDD dataset has 1,25,973 patterns, whereas the testing dataset contains 22,544 patterns.

B. Evaluating the proposed combined feature selection methodology (CFSM)

Many feature selection techniques are investigated to the recommended CFSM based on NB classifier to demonstrate the efficacy of the proposed CFSM. The following are the most recent feature selection techniques for investigation: vote scheme and information gain (IG) [27], cuttlefish algorithm (CFA) [28], highest wins (HW) algorithm [29], and particle swarm optimization (PSO) [30]. Considering NSL-KDD dataset, the selected features from each feature selection technique are shown in Table III. As stated by Table IV, the accuracy, precision, recall, and F1-score for CFSM is 96%, 94.8%, 95% and 94.9% respectively. Therefore, CFSM is superior to IG, CFA, HW, and PSO. The fundamental reason for the suggested CFSM method's superior performance is that it merges the advantages of the filter as well as wrapper techniques. CFSM picks a more relevant as well as powerful features from the input dataset, allowing the attack to be differentiated from regular dataset occurrences.

TABLE III
SELECTED SET OF FEATURES FROM NSL-KDD BY DIFFERENT FEATURE SELECTION ALGORITHMS

Technique	Features	Selected set of features
IG	8	[5,3,6,4,30,29,33,24]
PSO	37	[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]
CFA	10	[4,10,13,22,23,24,29,35,36,41]
HW	8	[4,5.6,12,28,30,31,35]
CFSM	18	[1,2,3,4,5,9,11,20,26,28,29,30,31,32,33,35,36,37]

TABLE IV
RESULTS OF DIFFERENT FEATURE SELECTION TECHNIQUES USING NSL-KDD DATA SET

Technique	accuracy	recall	precision	F1-score	Run time(s)
CFA	90%	82%	81.7%	81.8%	16
PSO	78.2%	77%	76%	76.5%	13
IG	80%	70%	75%	72.4%	12
HW	88.3%	88.3%	88.6%	88.2%	10
CFSM	96%	95%	94.8%	94.9%	8

C. Evaluating the proposed prioritized naïve bayes phase (PNBP)

The proposed Prioritized Naive Bayes Phase (PNBP) will be examined in this section. The PNBP is compared to the latest classification techniques, which include; extreme gradient-boosting (XGBoost) [31], gradient boosted decision tree (GBDT) [32], and particle swarm optimization-based

probabilistic neural network (PSO-PNN) [33]. According to Table V, the accuracy, precision, recall, and F1-score for PNBP are 97%, 98.6%, 98.5% and 98.4% respectively. Therefore, the performance of PNBP is much better and faster than XGBoost, GBDT, and PSO-PNN.

D.Evaluating the proposed Intrusion Detection Strategy (IDS)

To demonstrate the efficacy of our suggested IDS, it is compared to some of the most commonly intrusion detection techniques such as; Differential Evolution-Extreme Learning Machine (DE-ELM) [15], Adaptive Sampling-Convolutional Neural Network (AS-CNN) [14], Genetic Algorithm-Artificial Neural Network (GA-ANN) [12], and Bidirectional Attention Mechanism-Multiple Convolutional (BAT-MC) [13]. Results are shown in Table VI. It is noted that the IDS has competitive performance than DE-ELM, AS-CNN, GA-ANN, and BAT-MC. Because, the proposed phases in CS; PNBP and DEP are based on the essential features for intrusion detection that are picked through FSS, IDS provides speedy and exact recognition for the intrusion attempt.

TABLE V
RESULTS OF PNBP AND THE OTHER CLASSIFICATION APPROACHES

Technique	Accuracy	Recall	Precision	F1-score	Run time(s)
XGBoost	95.5%	98%	92%	95%	17
GBDT	86.10%	78.48%	96.44%	86.54%	14
PSO-PNN	95%	95.5%	97%	96.2%	15
PNBP	97%	98.5%	98.6%	98.4%	10

TABLE VI
COMPARISON BETWEEN IDS AND THE RECENTLY INTRUSION DETECTION APPROACHES

Technique	Accuracy	Recall	Precision	F1-score
DE-ELM	87.53%	81%	80%	80.5%
AS-CNN	80%	75%	74%	74.4%
GA-ANN	83.2%	78%	77%	77.5%
BAT-MC	84.3%	80%	79%	79.5%
Proposed IDS	97.6%	98.24%	98.14%	98.11%

V. CONCLUSIONS

The recommended Intrusion Detection Strategy (IDS) is comprised of three main steps: (i) Preparing Step (PS), (ii) Feature Selection Step (FSS), and (iii) Classification Step (CS). PS monitors and analyzes network activity in order to generate data for training and testing. The suggested Combined Feature Selection methodology (CFSM) incorporates the advantages of either filter and wrapper selection approaches. CFSM chooses useful as well as informative features from PS. The chosen features are then prioritized to power the proposed classification system, which has two phases named PNBP and DEP to make consistent and comparable judgments. When compared to other contemporary techniques utilizing the NSL-KDD dataset, the evaluation results revealed that the suggested IDS gives rapid as well as precise outcomes for Accuracy, Precision, Recall, F1-measure, and Run Time. In the future, the author plans to test edge computing and machine learning algorithms in a real-time situation to see if they can identify intrusion attacks efficiently. Furthermore, this study invites further research into the applicability of the proposed approach to other datasets.

FUNDING STATEMENT:
No funds, grants, or other support was received.

DECLARATION OF CONFLICTING INTERESTS STATEMENT:

The author has no conflicts of interest to declare that are relevant to the content of this article.

REFERENCES

[1] L. Thi-Thu-Huong, K. Yongsu, K. Howon. (2019, April). Network Intrusion Detection Based on Novel Feature Selection Model and Various Recurrent Neural Networks. Appl. Sci. 9(7):1392. Available: https://doi.org/10.3390/app9071392

[2] S. Samira, S.N.F. Mohd, H.Z. Mohd, A.M. Taufik. (2020, April). An Efficient Anomaly Intrusion Detection Method with Feature Selection and Evolutionary Neural Network. IEEE Access. 8, pp.70651–70663. Available: DOI: 10.1109/ACCESS.2020.2986217

[3] I. Ullah, Q.H. Mahmoud. (2017, December). A Filter-based Feature Selection Model for Anomaly-based Intrusion Detection Systems. Presented at Proceedings of the 2017 IEEE International Conference on Big Data. Available: DOI: 10.1109/BigData.2017.8258163.

[4] P. Mishra, V. Varadharajan, U. Tupakula, E.S. Pilli. (2019, June). A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection. IEEE Communications Surveys & Tutorials. 21, pp.686–728. Available: DOI: 10.1109/COMST.2018.2847722

[5] S. Fong, G.Li, N. Dey, R. Crespo, E. Viedma. (2020). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl. Soft Comput., 93, pp. 1–14. Available: https://doi.org/10.1016/j.asoc.2020.106282

[6] J. Suri, A. Puvvula, M. Biswas. (2020, August). COVID-19 pathways for brain and heart injury in comorbidity patients: a role of medical imaging and artificial intelligence-based COVID severity classification: a review. Comput. Biol. Med. 124, pp. 1–15. Availabel: doi: 10.1016/j.compbiomed.2020.103960

[7] N. Harzevili, S. Alizadeh. (2018, August). Mixture of latent multinomial naïve Bayes classifier. Appl. Soft Comput. 69, pp. 516–567. Available: https://doi.org/10.1016/j.asoc.2018.04.020

[8] Y. Kwon, A. Kwasinski, and A. Kwasinski. (2019). Solar irradiance forecast using Naïve Bayes classifier based on publicly available weather forecasting variables. Energies. 12(8), pp. 1–13. Available: Doi: 10.3390/en12081529.

[9] M. Sharad, J. Bhavesh. (2019). Application of Naïve Bayes classification for disease prediction. Int. J. Manag. 9(4), pp. 80–87.

[10] J. Kolluri, S. Razia. (2020). Text classification using Naïve Bayes classifier. Materialstoday. Available: https://Doi.org/10.1016/j.matpr.2020.10.058

[11] H. Chen, S. Hu, R. Hua, X. Zhao. (2021). Improved naive Bayes classification algorithm for traffic risk management. EURASIP Journal on Advances in Signal Processing. 30, pp. 4149–4162. Available: https://Doi.org/10.1186/s13634-021-00742-6

[12] S. Hosseini. (2020, April). A new machine learning method consisting of GA-LR and ANN for attack detection. Wireless Networks. 26(6), pp. 4149–4162, 2020
Available: https://doi.org/10.1007/s11276-020-02321-3

[13] T. Su, H. Sun, J. Zhu, S. Wang and, Y. Li. (2020, February). BAT: deep learning methods on network intrusion detection using NSL-KDD

dataset. IEEE Access. 8, pp. 29575–29585. Available: 10.1109/ACCESS.2020.2972627
[14] Z. Hu, L. Wang, L. Qi, Y. Li and W. Yang. (2020, October). A Novel Wireless Network Intrusion Detection Method Based on Adaptive Synthetic Sampling and an Improved Convolutional Neural Network. IEEE Access. 8, pp. 195741–195751. Available: 10.1109/ACCESS.2020.3034015
[15] F.H. Almasoudy, W.L. Al-Yaseen, and A.K. Idrees. (2020). Differential Evolution Wrapper Feature Selection for Intrusion Detection System. Procedia Computer Science. 167, pp. 1230–1239. Available: https://doi.org/10.1016/j.procs.2020.03.438
[16] A. I. Saleh, A. I. El Desouky, S.H. Ali. (2015, February). Promoting the performance of vertical recommendation systems by applying new classification techniques. Knowledge-Based System. 75, pp. 192–223. Available: https://doi.org/10.1016/j.knosys.2014.12.002
[17] P. khare, K. Burse. (2016). Feature selection using genetic algorithm and classification using weka for Ovarian Cancer. Int. J. Comput. Sci. Inf. Technol. (IJCSIT). 7 (1), pp. 194–196.
[18] S. H. Ali, R. A. El- Atier, K. M. Abo- Al- Ez, A. I. Saleh. (2020, April). A Gen- Fuzzy Based Strategy (GFBS) for Web Service Classification. Wireless Personal Communications. 113, pp. 1917–1953. Available: https://doi.org/10.1007/s11277-020-07300-7
[19] M. I. Prasetiyowati, N. U. Maulidevi, K. Surendro. (2021, June). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. Prasetiyowati et al. J Big Data. 8(84). Available: https://doi.org/10.1186/s40537-021-00472-4
[20] S. Bahassine, A. Madani, M. Al-Sarem, M. Kissi. (2020). Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University – Computer and Information Sciences.32, pp. 225-231. Available: https://doi.org/10.1016/j.jksuci.2018.05.010
[21] H. Djellali, N. Zine, N. Azizi. (2016). Two stages feature selection based on filter ranking methods and SVMRFE on medical applications. Modelling and Implementation of Complex Systems Lecture Notes in Networks and Systems. 1, pp. 281–293.
[22] A. Wosiak, D. Zakrzewska. (2018). Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis. Complexity. Available: https://doi.org/10.1155/2018/2520706
[23] L. Jiang, Ch. Qiu, Ch. Li. (2015). A novel minority cloning technique for cost-sensitive learning. Int. J. Pattern Recognition and Artificial Intelligence. 29 (4), pp. 1–18. Available: https://doi.org/10.1142/S0218001415510040
[24] A. Rabie, S. Ali, A. Saleh, H. Ali. (2020). A fog based load forecasting strategy based on multi-ensemble classification for smart grids. J. Ambient Intell. Humanized Comput. 11 (1), pp. 209–236.
[25] M. Ahmed, A. N. Mahmood, and J. Hu. (2016, January). A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 60, pp. 19–31.
[26] S. Revathi, and A. Malathi. (2013). A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology (IJERT). 2 (12), pp. 1848–1853.
[27] S. Aljawarneh, M. Aldwairi, and M. B. Yassein. (2018, March). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science. 25, pp. 152–160. Available: https://doi.org/10.1016/j.jocs.2017.03.006
[28] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour. (2019, February). Cyber intrusion detection by combined feature selection algorithm. Journal of Information Security and Applications, 44, pp. 80–88. Available: https://doi.org/10.1016/j.jisa.2018.11.007
[29] R. M. A. Mohammad, M. K. Alsmadi. (2021). Intrusion detection using Highest Wins feature selection algorithm. Neural Computing and Applications. 33, pp. 9805–9816. Available: https://doi.org/10.1007/s00521-021-05745-w
[30] B. A. Tama, M. Comuzzi, and, K.-H. Rhee. (2019, July). TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System. IEEE Access.7, pp. 94497–94507. Available: DOI: 10.1109/ACCESS.2019.2928048
[31] A.O. Alzahrani, M.J.F. Alenazi. (2021, April). Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks. Future Internet. 13(5). Available: https://doi.org/10.3390/fi13050111
[32] L. Li, Y. Yu, S. Bai, J. Cheng, and X. Chen. (2018). Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO. Journal of sensors. 6, pp. 1-9. Available: https://DOI:10.1155/2018/1578314
[33] M. Rabbani, Y. L. Wang, R. Khoshkangini, H. Jelodar, R. Zhao, P. Hu. (2020, February). A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing. Journal of Network and Computer Applications. 151. Available: https://DOI:10.1016/j.jnca.2019.102507

Title Arabic:

استراتيجية جديدة لكشف التسلل تعتمد على منهجية اختيار الميزات المجمعة وتقنية التعلم الآلي

Arabic Abstract:

نظام كشف التطفل هو آلية أمان مهمة تراقب حركة مرور الشبكة للمساعدة في منع الوصول غير المرغوب فيه إلى موارد الشبكة. يعد الكشف الفعال عن التسلل مسألة مهمة للدفاع عن الشبكات ضد الاختراقات المحتملة. في هذه الورقة ، تم اقتراح إستراتيجية جديدة لكشف التسلل .(IDS) تنقسم أنظمة تحديد الهوية (IDS)الموصى بها إلى ثلاث خطوات: (١) خطوة التحضير (PS) ، (2) خطوة اختيار الميزة (FSS) ، وخطوة التصنيف .(CS) يجمع PS ويحلل حركة مرور الشبكة استعدادًا للتدريب والاختبار. يهدف FS^2 إلى اختيار الميزات المهمة لاكتشاف هجمات التسلل من .PS وهي تتألف من وحدتين متتاليتين لاختيار الميزات ، وهما ؛ وحدة الاختيار السريع (QSM) ووحدة التحديد الدقيق .(PSM)تستخدم PSM الخوارزمية الجينية (GA) كطريقة مجمعة ، بينما تعتمد QSM على عامل التصفية. استنادًا إلى الميزات الأكثر فاعلية التي حددتها FS^2 ، يسعى CS إلى اكتشاف هجمات التطفل بأقل قدر من العقوبة الزمنية. يحتوي على مرحلتين: مرحلة Naive Bayes ذات الأولوية (PNBP) ومرحلة تشجيع المسافة (DEP) ، والتي تتجنب مشاكل مصنفات Naive Bayes النموذجية .(NB) يتفوق نظام IDS الموصى به على الأساليب السابقة الأخرى باستخدام مجموعة بيانات NSL-KDD ، وفقًا للاختبارات التجريبية. يوفر نظام IDSأعلى دقة ودقة واسترجاع وقياس F1 بقيم تساوي ٩٧.٦٪ و ٩٨.٢٤٪ و ٩٨.١٤٪ و ٩٨.١١٪ على التوالي مع الحد الأدنى من عقوبة الوقت